School of Information and Library Science
University of North Carolina at Chapel Hill
October 27, 2003

From *terminology* integration to *information* integration

An example in the domain of genetics

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Outline

- Background
  - Terminology integration:
    The Unified Medical Language System
  - Information integration:
    Genetics as an example
- Applications
  - GenesTrace
  - BioMeKe

Terminology integration

The Unified Medical Language System

Motivation

- Started in 1986
- National Library of Medicine
- “Long-term R&D project”
- Complementary to IAIMS

(Integrated Academic Information Management Systems)

«[…] the UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information.

- The first is the variety of ways the same concepts are expressed in different machine-readable sources and by different people.
- The second is the distribution of useful information among many disparate databases and systems.»

Source Vocabularies

- 117 “sources”
- ~60 families of vocabularies
  - multiple translations (e.g., MeSH, ICPC, ICD-10)
  - variants (American-English equivalents, Australian extension/adaptation)
  - subsequent versions usually considered distinct families (ICD: 9-10; DSM: IIIR-IV)
- Broad coverage of biomedicine
- Common presentation

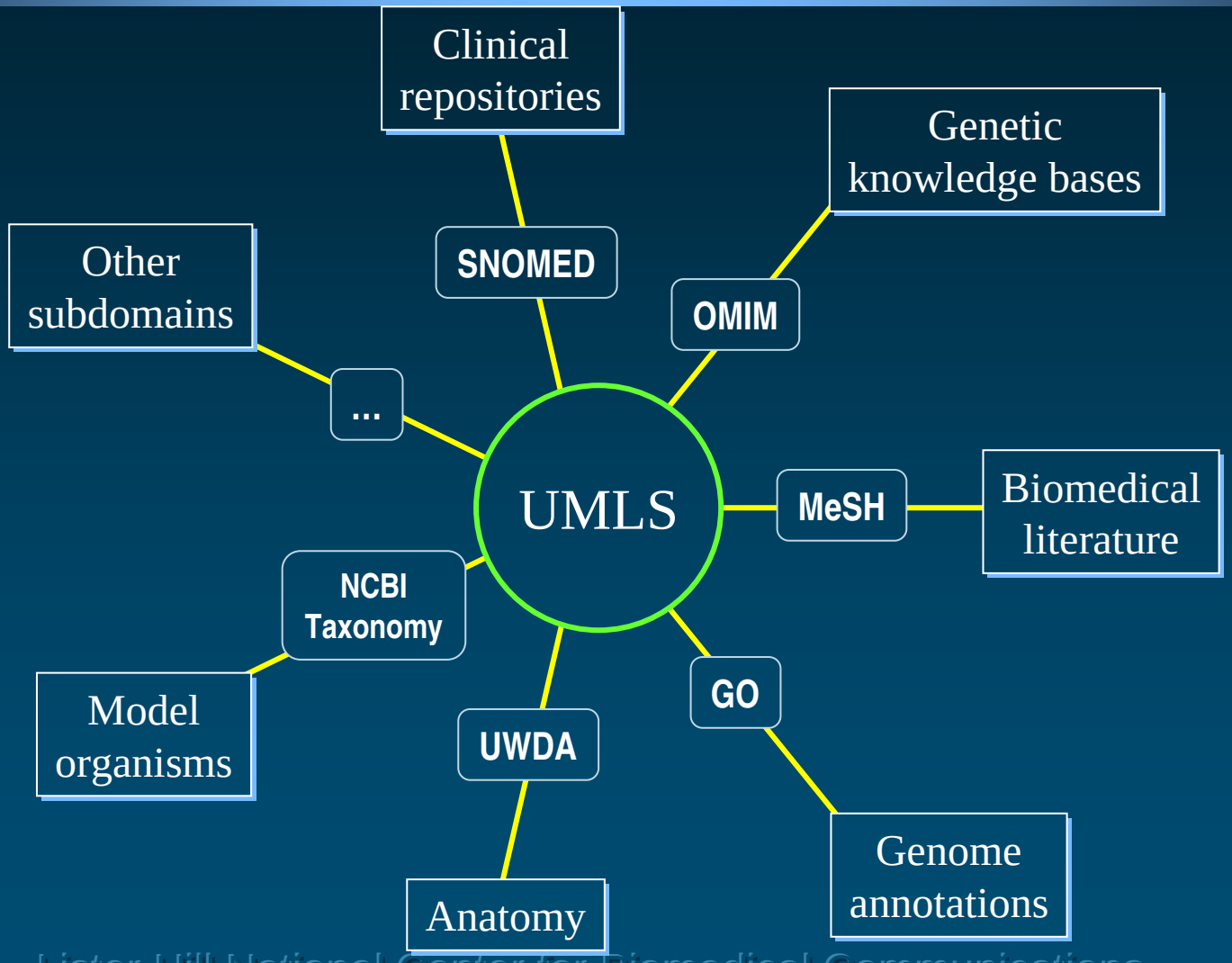
Biomedical terminologies

- Core vocabularies
  - anatomy (UWDA, Neuronames)
  - drugs (First DataBank, Micromedex)
  - medical devices (UMD, SPN)
- Several perspectives
  - clinical terms (SNOMED, CTV3)
  - information sciences (MeSH, CRISP)
  - administrative terminologies (ICD-9-CM, CPT-4)
  - standards (HL7, LOINC)

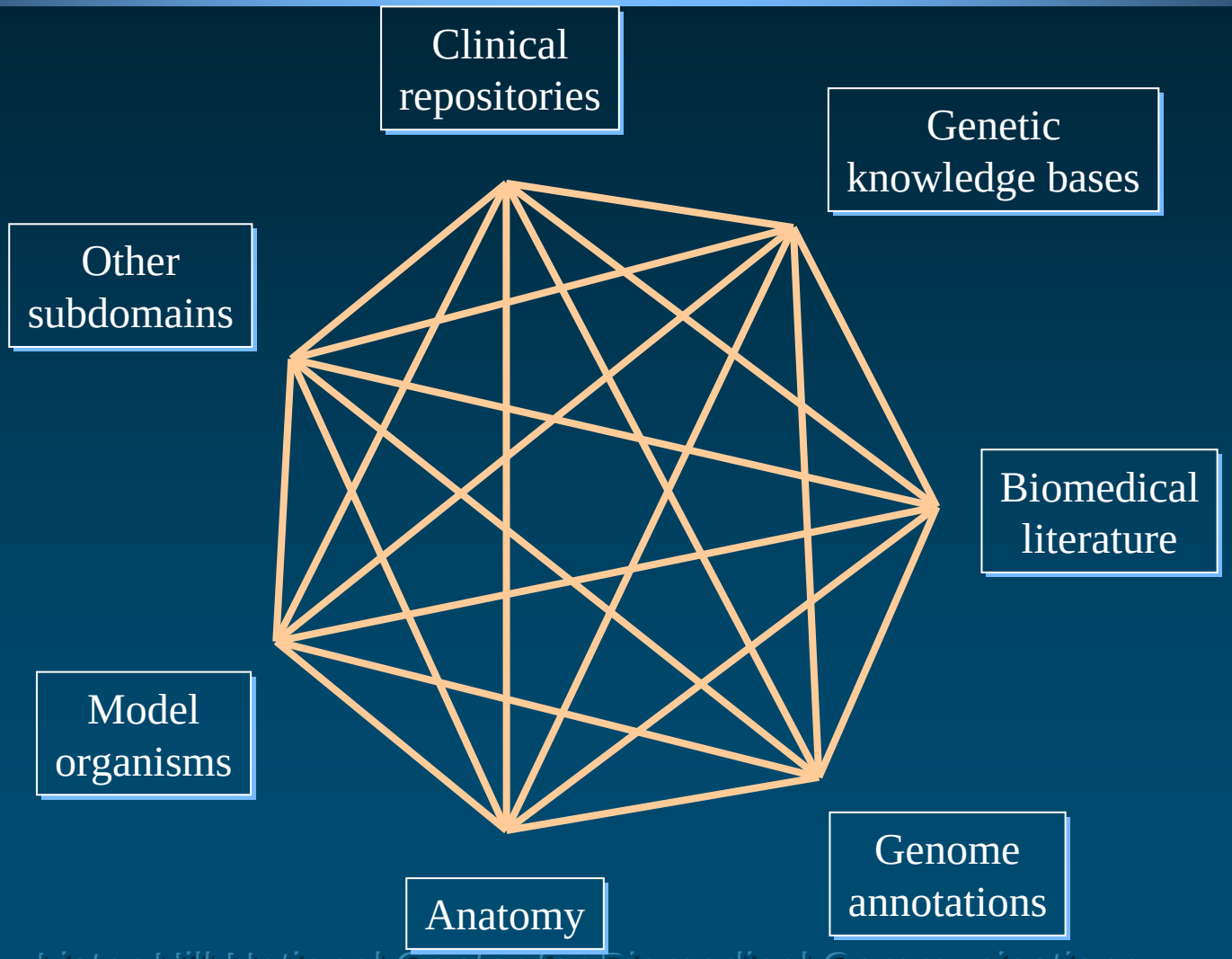
Biomedical terminologies (cont'd)

- Specialized vocabularies
  - nursing (NIC, NOC, NANDA, Omaha, PCDS)
  - dentistry (CDT)
  - oncology (PDQ)
  - psychiatry (DSM, APA)
  - adverse reactions (COSTART, WHO ART)
  - primary care (ICPC)
- Knowledge bases (AI/Rheum, DXplain, QMR)

Integrating subdomains



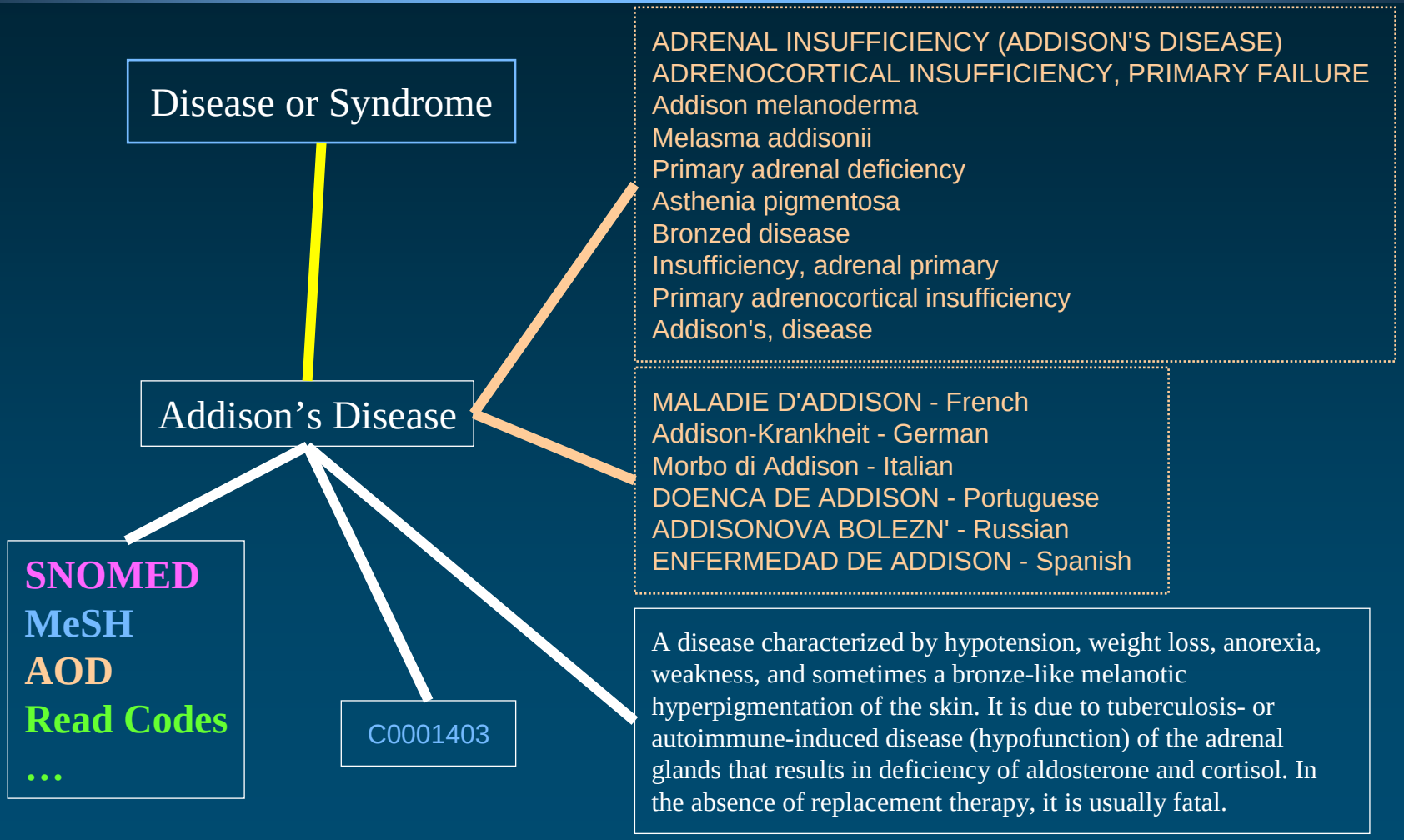
Integrating subdomains



UMLS: 3 components

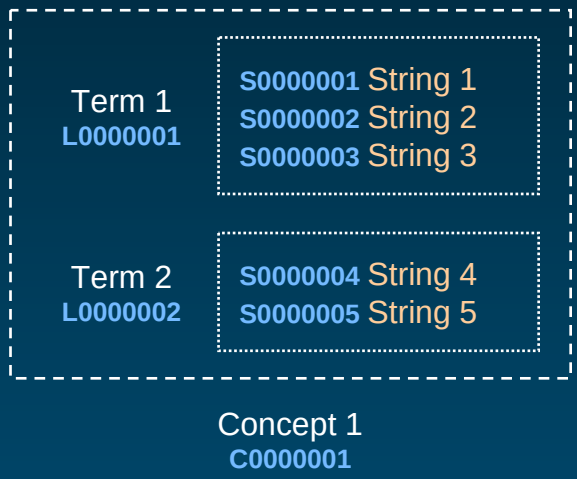
- Metathesaurus
  - Concepts
  - Inter-concept relationships
- Semantic Network
  - Semantic types
  - Semantic network relationships
- Lexical resources
  - SPECIALIST Lexicon
  - Lexical tools

Addison's Disease: Concept

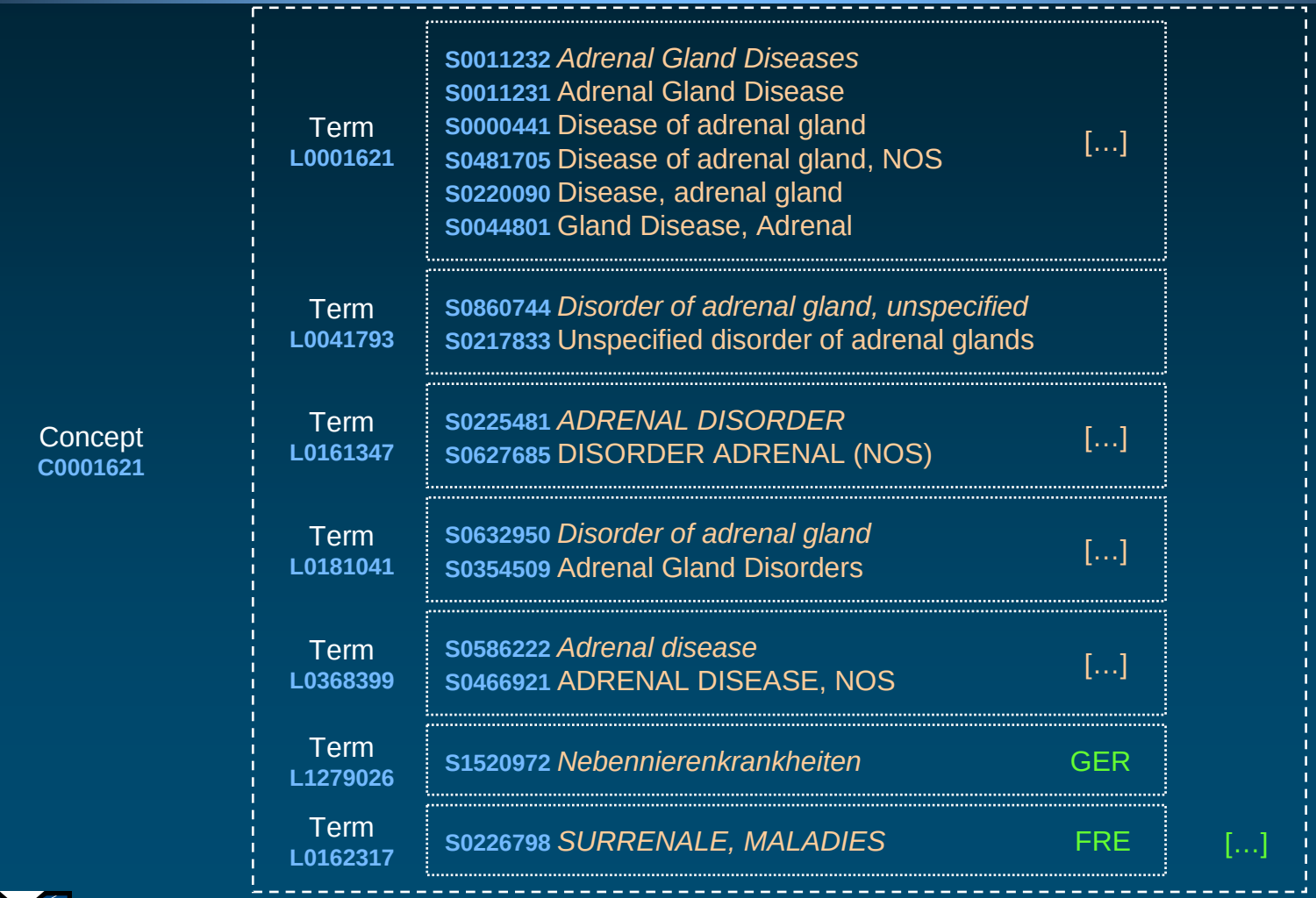


Metathesaurus Concepts (2003AA)

- Concept: Cluster of synonymous terms
  - ~875,000 concepts
  - identified by a CUI
- Term: Set of lexical variants
  - ~1.8 M terms
  - identified by a LUI
- String: Concept name
  - ~2.1 M strings
  - identified by a SUI



Cluster of synonymous terms



Metathesaurus Relationships

- Symbolic relations: ~5 M pairs of concepts
- Statistical relations : ~6.5 M pairs of concepts (co-occurring concepts)

- Categorization: Relationships between concepts and semantic types from the Semantic Network

Symbolic relations

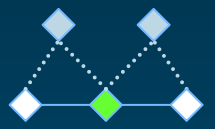
- Relation

  MRREL

  - Pair of concept identifiers
  - Type
  - Attribute (if any)
  - List of sources (for type and attribute)
- Semantics of the relationship: defined by its type [and attribute]

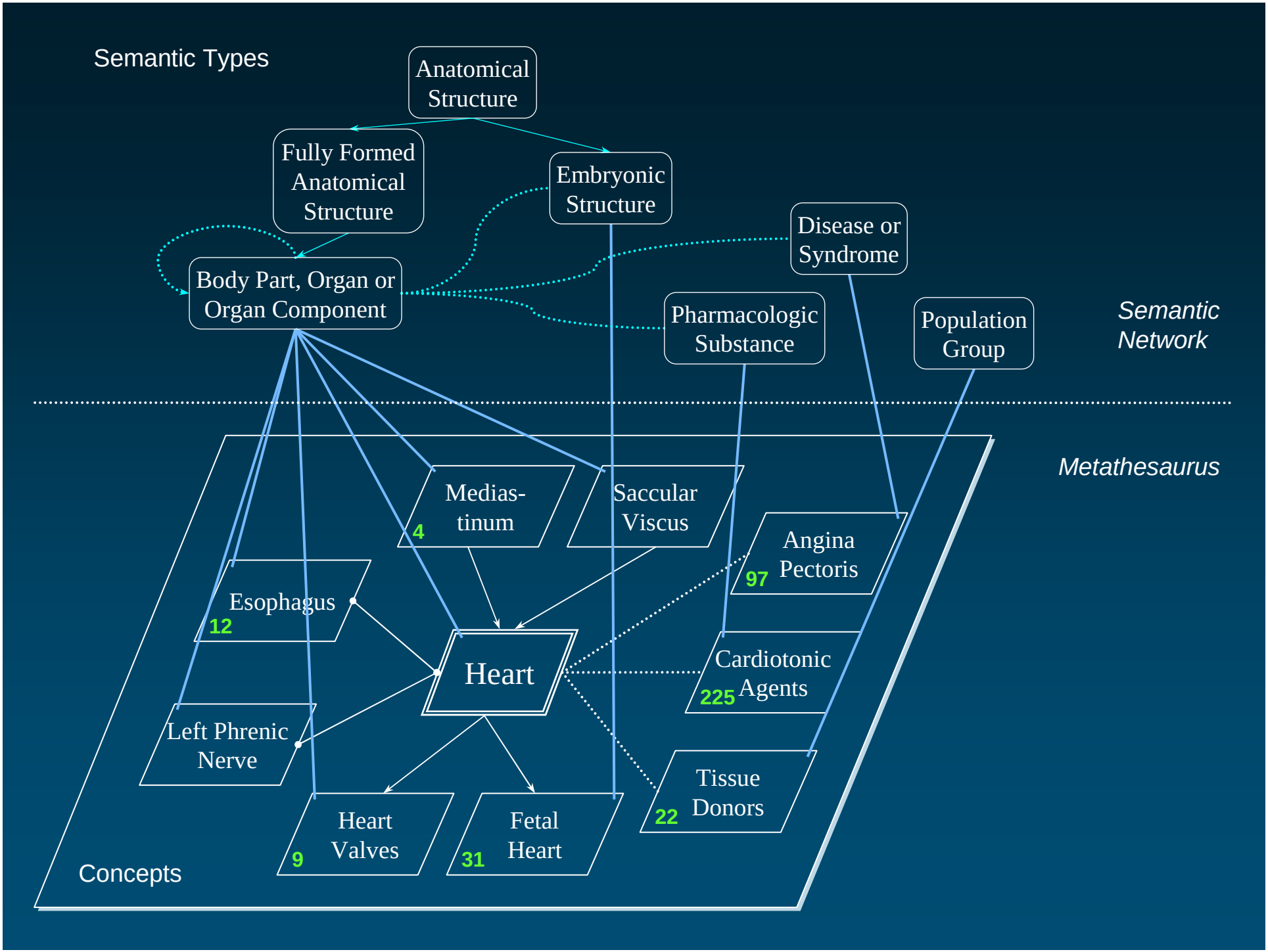
Symbolic relationships Type

- Hierarchical
  - Parent / Child — **PAR/CHD**
  - Broader / Narrower than — **RB/RN**
- Derived from hierarchies
  - Siblings (children of parents) — **SIB**
- Associative
  - Other — **RO**
- Various flavors of near-synonymy
  - Similar — **RL**
  - Source asserted synonymy — **SY**
  - Possible synonymy — **RQ**



Symbolic relationships Attribute

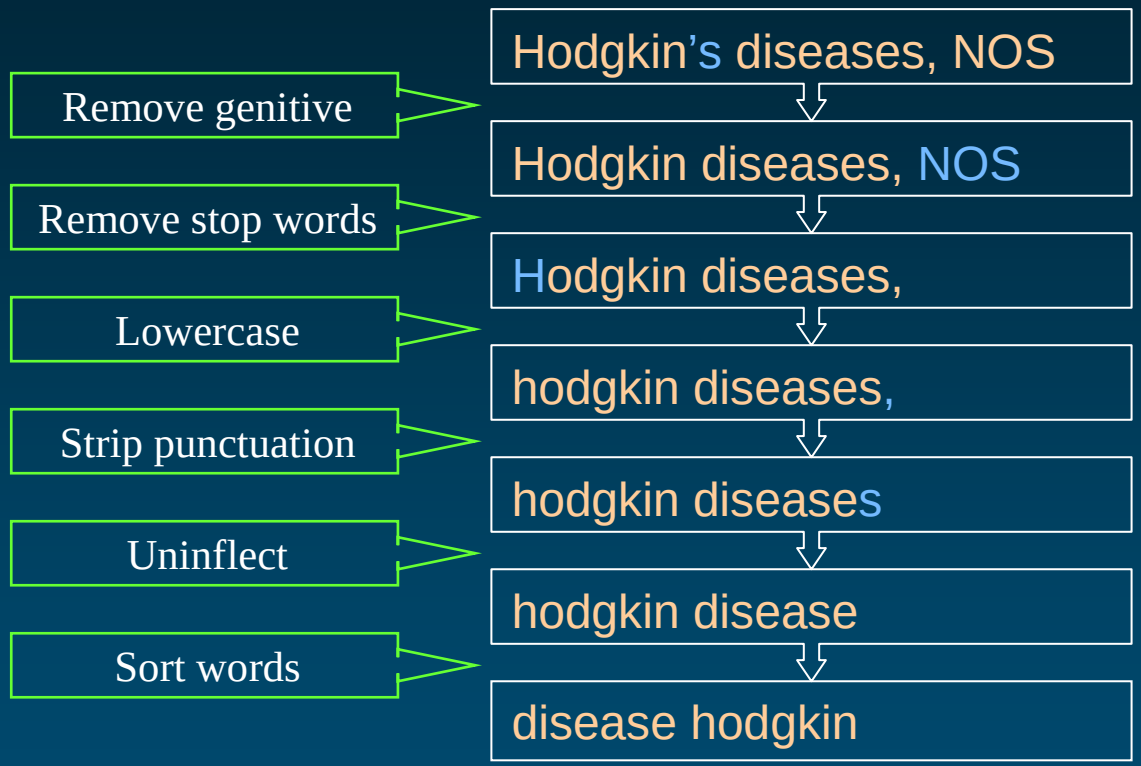
- Hierarchical
  - isa (is-a-kind-of)
  - part-of
- Associative
  - location-of
  - caused-by
  - treats
  - …
- Cross-references (mapping)


Semantic Types
Anatomical Structure
Fully Formed Anatomical Structure
Embryonic Structure
Disease or Syndrome
Body Part, Organ or Organ Component
Pharmacologic Substance
Population Group
Semantic Network
Metathesaurus
Medias-tinum
4
Saccular Viscus
Angina Pectoris
97
Esophagus
12
Heart
Cardiotonic Agents
225
Left Phrenic Nerve
Tissue Donors
22
Heart Valves
9
Fetal Heart
31
Concepts

Lexical tools

- To manage lexical variation in biomedical terminologies
- Major tools
  - Normalization
  - Indexes
  - Lexical Variant Generation program (lvg)
- Based on the SPECIALIST Lexicon
- Used by noun phrase extractors, search engines

Normalization



Normalization: Example

Hodgkin Disease
HODGKINS DISEASE
Hodgkin's Disease
Disease, Hodgkin's
Hodgkin's, disease
HODGKIN'S DISEASE
Hodgkin's disease
Hodgkins Disease
Hodgkin's disease NOS
Hodgkin's disease, NOS
Disease, Hodgkins
Diseases, Hodgkins
Hodgkins Diseases
Hodgkins disease
hodgkin's disease
Disease, Hodgkin

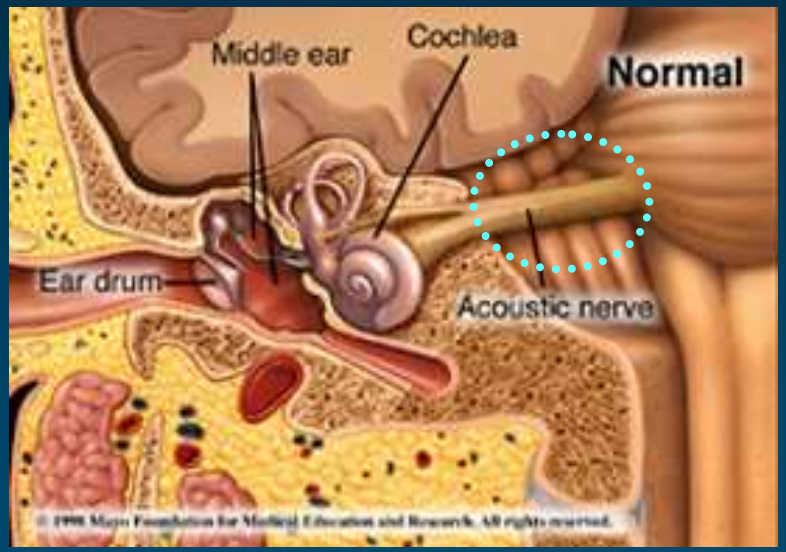
Information integration

Genetics as an example

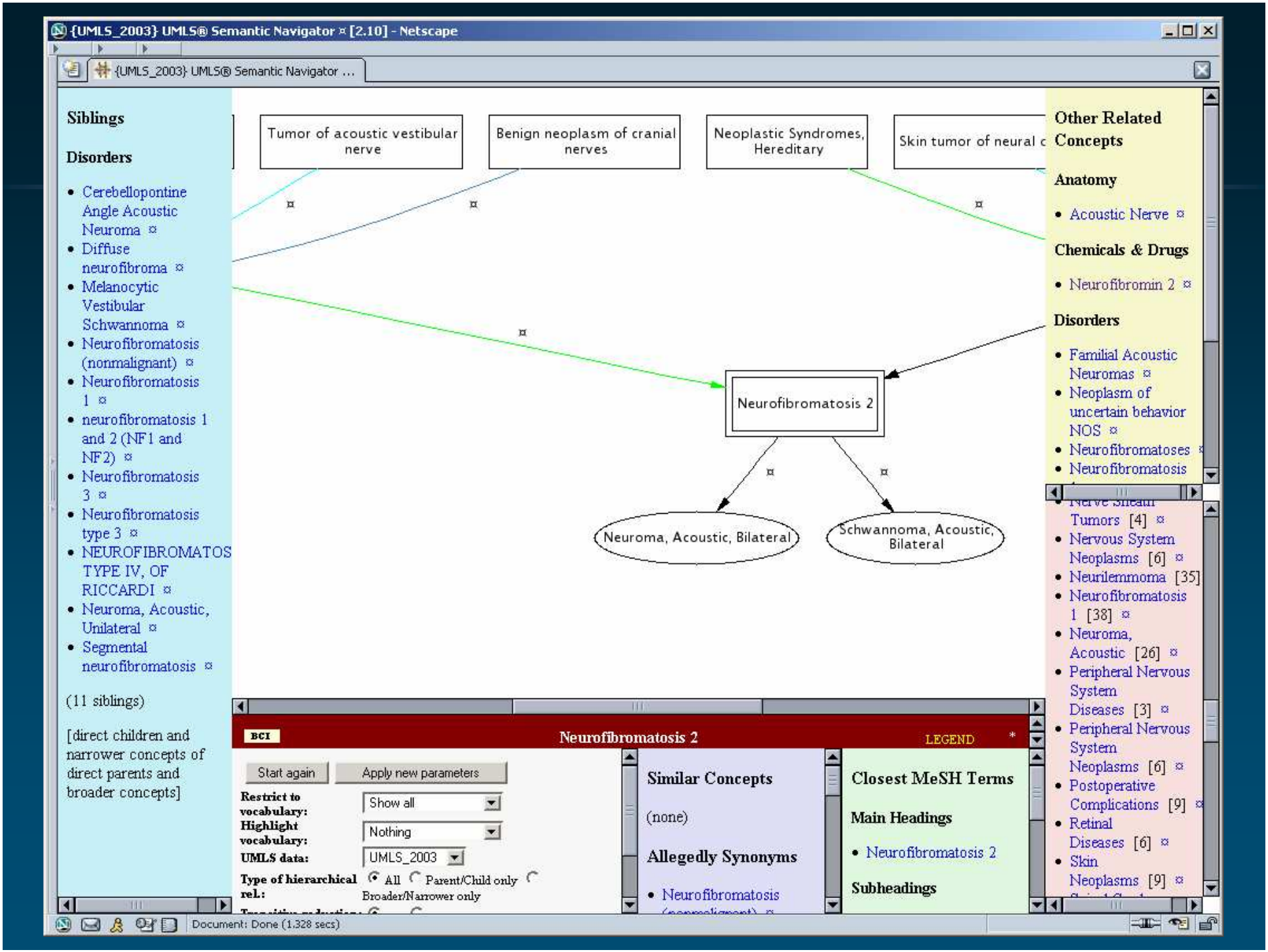
NF2 Gene, protein, and disease

Neurofibromatosis 2 is an autosomal dominant disease characterized by tumors called **schwannomas** involving the acoustic nerve, as well as other features. The disorder is caused by mutations of the **NF2** gene resulting in absence or inactivation of the protein product. The protein product of NF2 is commonly called **merlin** (but also **neurofibromin 2** and **schwannomin**) and functions as a tumor suppressor.

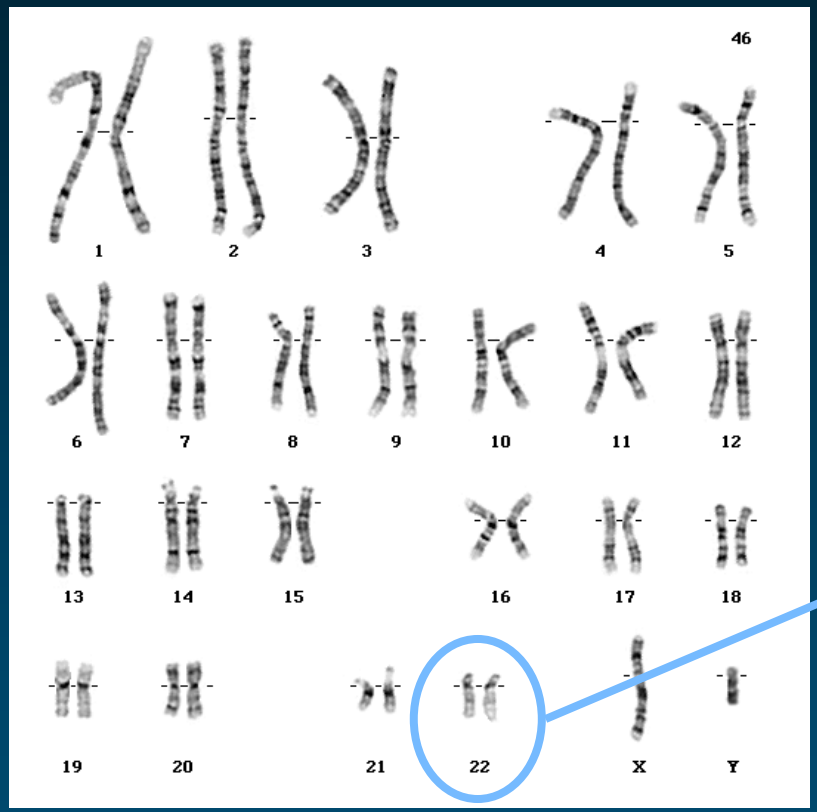
Schwannoma (acoustic neuroma)



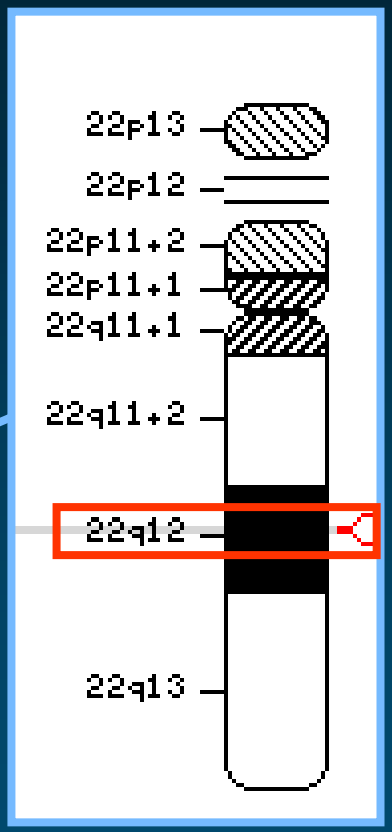
http://www.mayoclinic.com


{UMLS_2003} UMLS® Semantic Navigator ¤ [2.10] - Netscape
{UMLS_2003} UMLS® Semantic Navigator ...
Siblings
Disorders
Cerebellopontine Angle Acoustic Neuroma ¤
Diffuse neurofibroma ¤
Melanocytic Vestibular Schwannoma ¤
Neurofibromatosis (nonmalignant) ¤
Neurofibromatosis 1 ¤
neurofibromatosis 1 and 2 (NF1 and NF2) ¤
Neurofibromatosis 3 ¤
Neurofibromatosis type 3 ¤
NEUROFIBROMATOSIS TYPE IV, OF RICCARDI ¤
Neuroma, Acoustic, Unilateral ¤
Segmental neurofibromatosis ¤
(11 siblings)
[direct children and narrower concepts of direct parents and broader concepts]
Tumor of acoustic vestibular nerve
Benign neoplasm of cranial nerves
Neoplastic Syndromes, Hereditary
Skin tumor of neural c
Neurofibromatosis 2
Neuroma, Acoustic, Bilateral
Schwannoma, Acoustic, Bilateral
Other Related Concepts
Anatomy
Acoustic Nerve ¤
Chemicals & Drugs
Neurofibromin 2 ¤
Disorders
Familial Acoustic Neuromas ¤
Neoplasm of uncertain behavior NOS ¤
Neurofibromatoses ¤
Neurofibromatosis
Tumors [4] ¤
Nervous System Neoplasms [6] ¤
Neurilemmoma [35]
Neurofibromatosis 1 [38] ¤
Neuroma, Acoustic [26] ¤
Peripheral Nervous System Diseases [3] ¤
Peripheral Nervous System Neoplasms [6] ¤
Postoperative Complications [9] ¤
Retinal Diseases [6] ¤
Skin Neoplasms [9] ¤
BCI
Neurofibromatosis 2
LEGEND
Start again
Apply new parameters
Restrict to vocabulary:
Show all
Highlight vocabulary:
Nothing
UMLS data:
UMLS_2003
Type of hierarchical rel.:
All
Parent/Child only
Broader/Narrower only
Similar Concepts
(none)
Allegedly Synonyms
Neurofibromatosis
Closest MeSH Terms
Main Headings
Neurofibromatosis 2
Subheadings
Document: Done (1.328 secs)

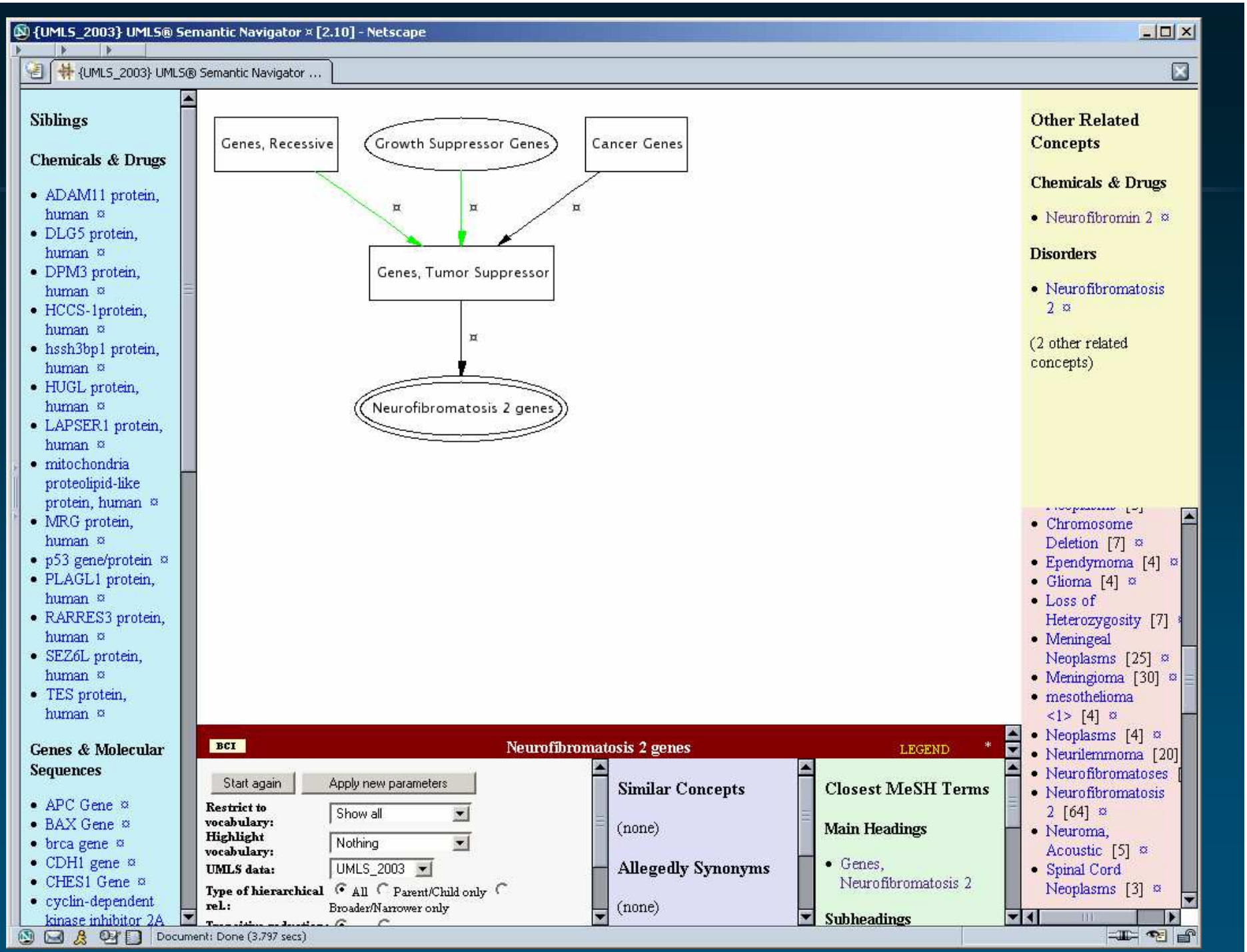
NF2 gene



http://staff.washington.edu/timk/cyto/human/

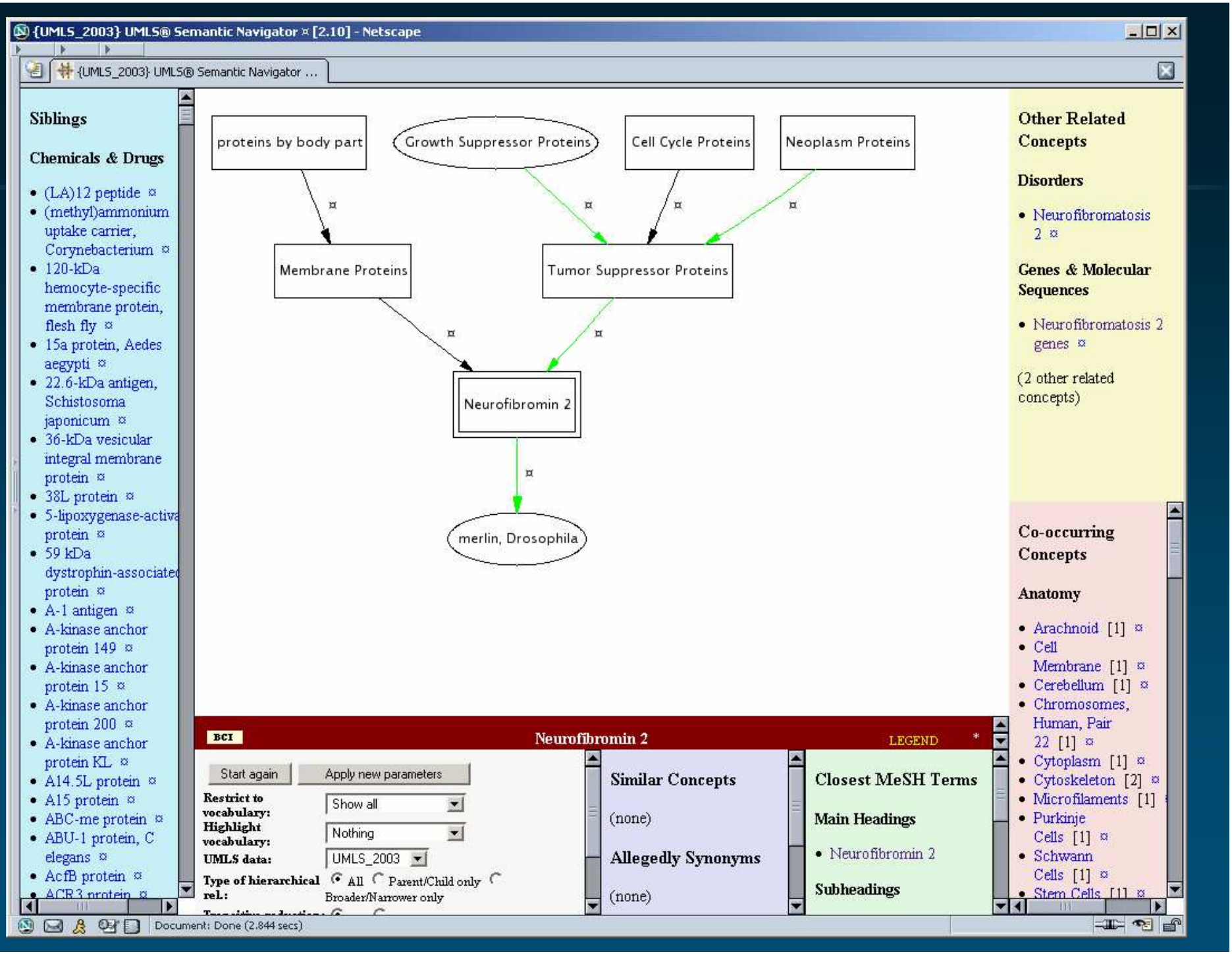


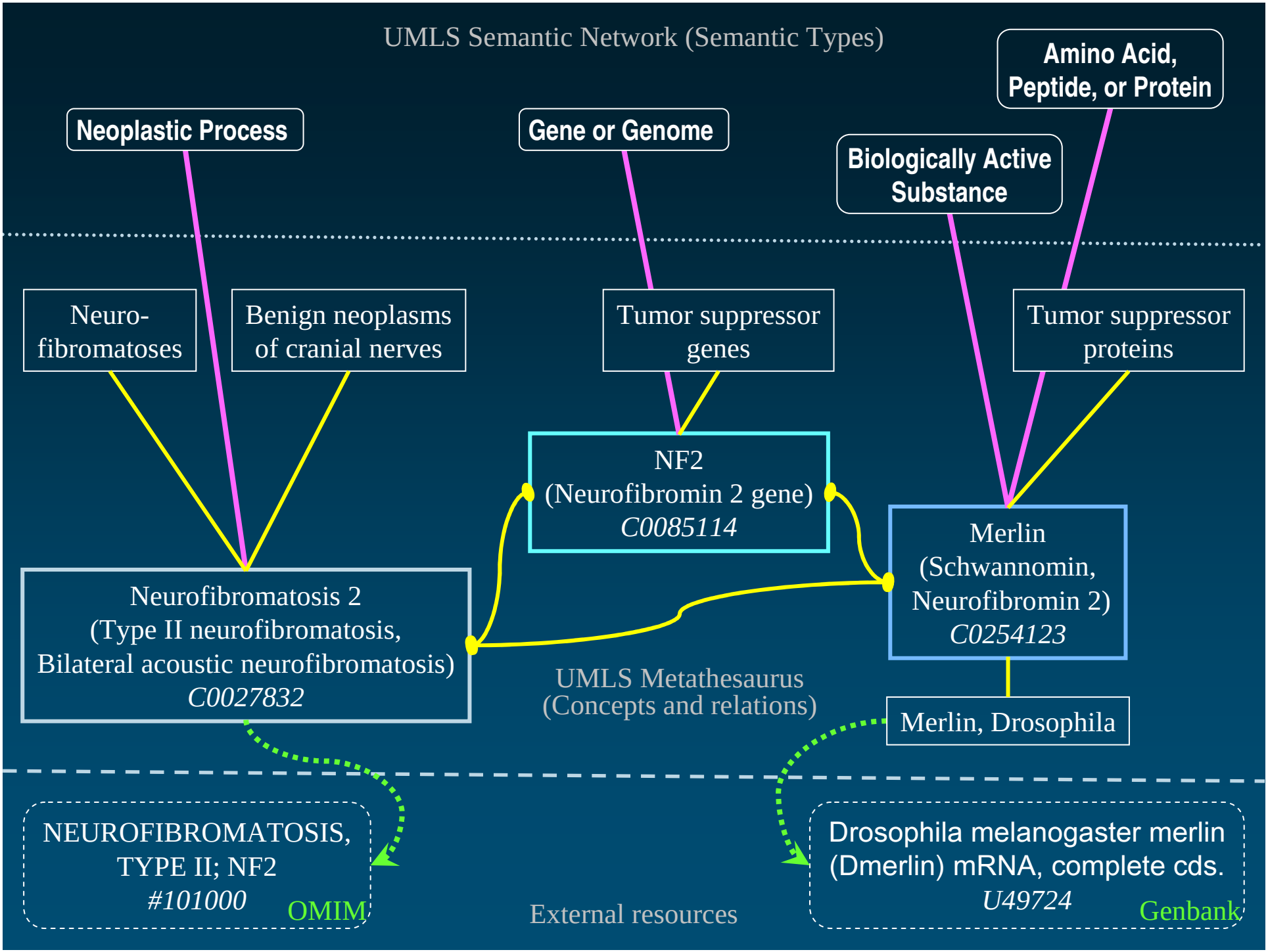
http://www.ncbi.nlm.nih.gov/mapview/


{UMLS_2003} UMLS® Semantic Navigator ¤ [2.10] - Netscape
{UMLS_2003} UMLS® Semantic Navigator ...
Siblings
Chemicals & Drugs
ADAM11 protein, human ¤
DLG5 protein, human ¤
DPM3 protein, human ¤
HCCS-1protein, human ¤
hssh3bp1 protein, human ¤
HUGL protein, human ¤
LAPSER1 protein, human ¤
mitochondria proteolipid-like protein, human ¤
MRG protein, human ¤
p53 gene/protein ¤
PLAGL1 protein, human ¤
RARRES3 protein, human ¤
SEZ6L protein, human ¤
TES protein, human ¤
Genes & Molecular Sequences
APC Gene ¤
BAX Gene ¤
brca gene ¤
CDH1 gene ¤
CHES1 Gene ¤
cyclin-dependent kinase inhibitor 2A
Genes, Recessive
Growth Suppressor Genes
Cancer Genes
Genes, Tumor Suppressor
Neurofibromatosis 2 genes
Other Related Concepts
Chemicals & Drugs
Neurofibromin 2 ¤
Disorders
Neurofibromatosis 2 ¤
(2 other related concepts)
Chromosome Deletion [7] ¤
Ependymoma [4] ¤
Glioma [4] ¤
Loss of Heterozygosity [7] ¤
Meningeal Neoplasms [25] ¤
Meningioma [30] ¤
mesothelioma <1> [4] ¤
Neoplasms [4] ¤
Neurilemmoma [20]
Neurofibromatoses
Neurofibromatosis 2 [64] ¤
Neuroma, Acoustic [5] ¤
Spinal Cord Neoplasms [3] ¤
BCI
Neurofibromatosis 2 genes
LEGEND
Start again
Apply new parameters
Restrict to vocabulary:
Show all
Highlight vocabulary:
Nothing
UMLS data:
UMLS_2003
Type of hierarchical rel.:
All
Parent/Child only
Broader/Narrower only
Similar Concepts
(none)
Allegedly Synonyms
(none)
Closest MeSH Terms
Main Headings
Genes, Neurofibromatosis 2
Subheadings
Document: Done (3.797 secs)

Merlin

- Synonyms
  - Neurofibromin 2
  - Schwannomin
  - Schwannomerlin
  - Neurofibromatosis-2
- 10 isoforms
- Annotations
  - Negative regulation of cell proliferation
  - Cytoskeleton
  - Plasma membrane


{UMLS_2003} UMLS® Semantic Navigator ¤ [2.10] - Netscape
{UMLS_2003} UMLS® Semantic Navigator ...
Siblings
Chemicals & Drugs
(LA)12 peptide ¤
(methyl)ammonium uptake carrier, Corynebacterium ¤
120-kDa hemocyte-specific membrane protein, flesh fly ¤
15a protein, Aedes aegypti ¤
22.6-kDa antigen, Schistosoma japonicum ¤
36-kDa vesicular integral membrane protein ¤
38L protein ¤
5-lipoxygenase-activa protein ¤
59 kDa dystrophin-associated protein ¤
A-1 antigen ¤
A-kinase anchor protein 149 ¤
A-kinase anchor protein 15 ¤
A-kinase anchor protein 200 ¤
A-kinase anchor protein KL ¤
A14.5L protein ¤
A15 protein ¤
ABC-me protein ¤
ABU-1 protein, C elegans ¤
AcfB protein ¤
ACR3 protein ¤
proteins by body part
Growth Suppressor Proteins
Cell Cycle Proteins
Neoplasm Proteins
Membrane Proteins
Tumor Suppressor Proteins
Neurofibromin 2
merlin, Drosophila
Other Related Concepts
Disorders
Neurofibromatosis 2 ¤
Genes & Molecular Sequences
Neurofibromatosis 2 genes ¤
(2 other related concepts)
Co-occurring Concepts
Anatomy
Arachnoid [1] ¤
Cell Membrane [1] ¤
Cerebellum [1] ¤
Chromosomes, Human, Pair 22 [1] ¤
Cytoplasm [1] ¤
Cytoskeleton [2] ¤
Microfilaments [1] ¤
Purkinje Cells [1] ¤
Schwann Cells [1] ¤
Stem Cells [1] ¤
BCI
Neurofibromin 2
LEGEND *
Start again
Apply new parameters
Restrict to vocabulary:
Show all
Highlight vocabulary:
Nothing
UMLS data:
UMLS_2003
Type of hierarchical rel.:
All
Parent/Child only
Broader/Narrower only
Similar Concepts
(none)
Allegedly Synonyms
(none)
Closest MeSH Terms
Main Headings
Neurofibromin 2
Subheadings
Document: Done (2.844 secs)


UMLS Semantic Network (Semantic Types)
Neoplastic Process
Gene or Genome
Biologically Active Substance
Amino Acid, Peptide, or Protein
Neuro-fibromatoses
Benign neoplasms of cranial nerves
Tumor suppressor genes
Tumor suppressor proteins
NF2 (Neurofibromin 2 gene) C0085114
Merlin (Schwannomin, Neurofibromin 2) C0254123
Neurofibromatosis 2 (Type II neurofibromatosis, Bilateral acoustic neurofibromatosis) C0027832
UMLS Metathesaurus (Concepts and relations)
Merlin, Drosophila
NEUROFIBROMATOSIS, TYPE II; NF2 #101000
OMIM
External resources
Drosophila melanogaster merlin (Dmerlin) mRNA, complete cds. U49724
Genbank

Limitations

- Genes not systematically represented
  - Most gene products and diseases are
- Gene/Gene product-Disease relations
  - Not systematically represented
  - Not explicitly represented (e.g., co-occurrence)
- Cross-references not systematically represented
- Naming conventions (genes)

Applications (1)

GenesTrace™

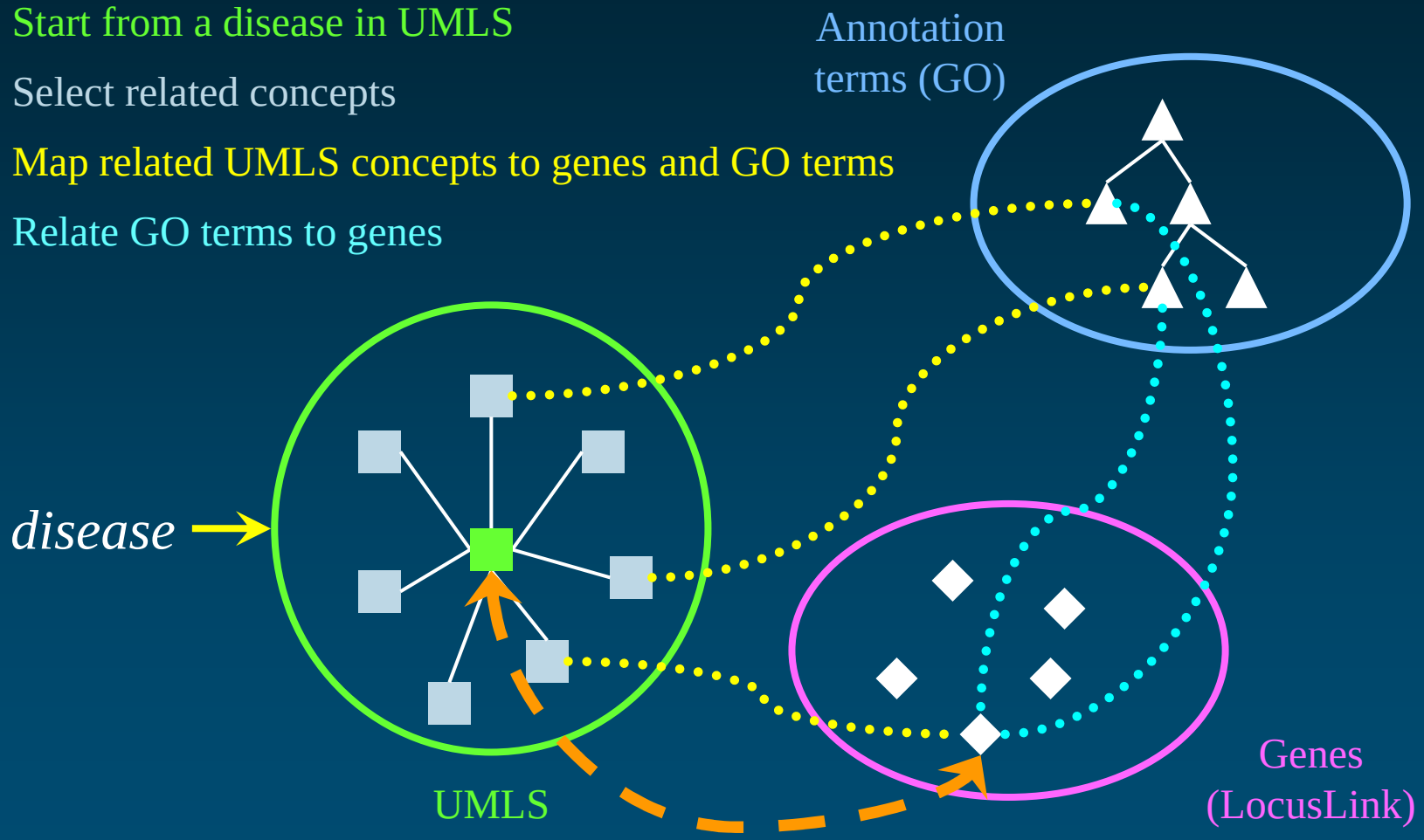
I.N. Sarkar & al.
Columbia University

Objectives

- Relate diseases to genes through structured, integrated terminologies
- Biological Knowledge Discovery

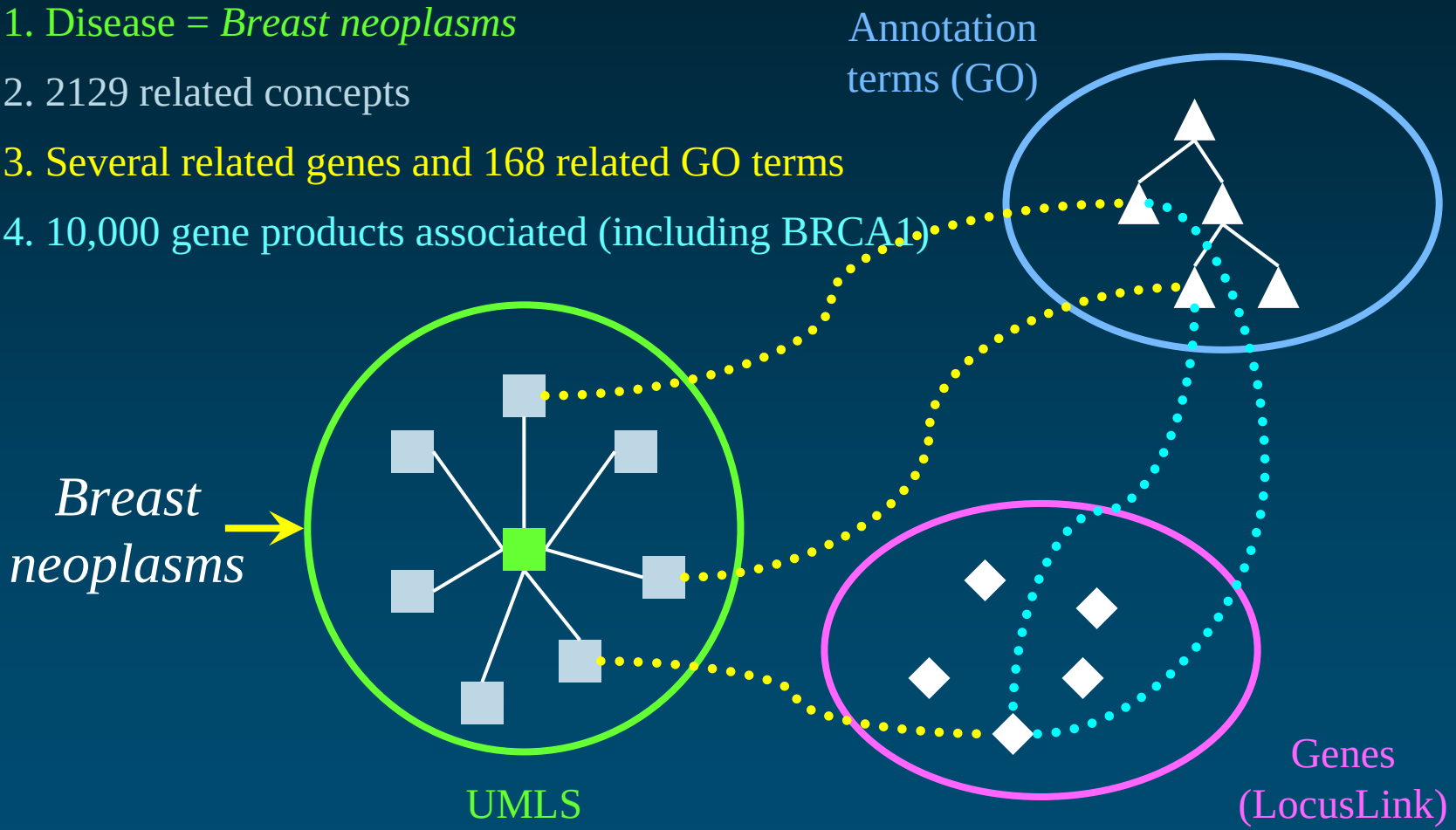
Resources and Methods

1. Start from a disease in UMLS
2. Select related concepts
3. Map related UMLS concepts to genes and GO terms
4. Relate GO terms to genes



Validation Breast cancer – BRCA1 association

1. Disease = *Breast neoplasms*
2. 2129 related concepts
3. Several related genes and 168 related GO terms
4. 10,000 gene products associated (including BRCA1)



Limitations

- Noise
  - Too many non-specific GO terms associated (e.g., *nucleus*)
  - Too many genes associated
- But
  - Promising preliminary results
  - Room for refinement

References

- I. Sarkar, M. Cantor, O. Bodenreider, Y. Lussier. GenesTrace™: Biological knowledge discovery via structured terminology – A feasibility study. Submitted to MEDINFO 2004.

Applications (2)

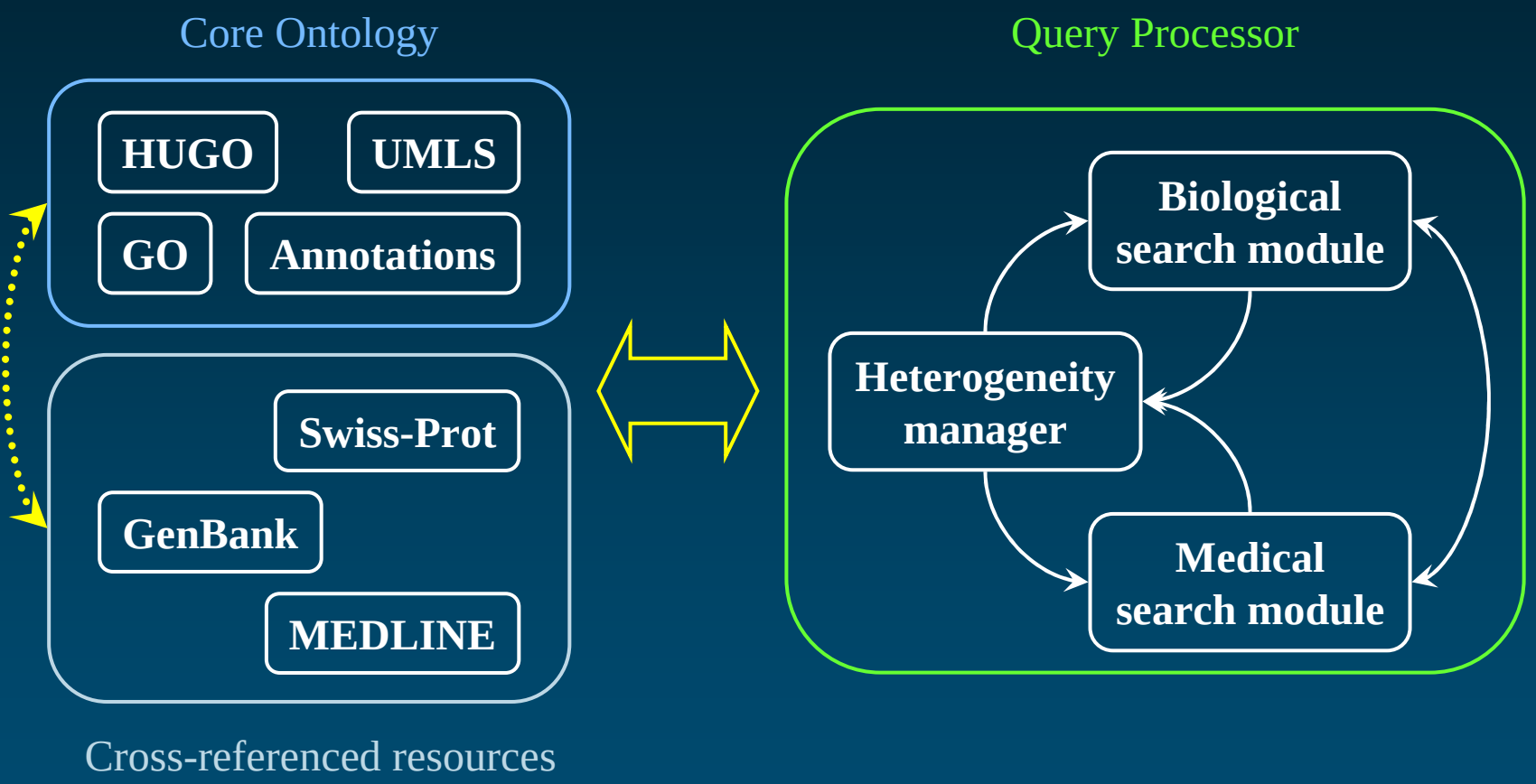
BioMeKe

G. Marquet & al.
LIM, Univ. Rennes, France

Objectives

- To develop a knowledge warehouse for transcriptome analysis (liver diseases)
- Semantic interoperability
  - Medical knowledge bases
  - Molecular biology and genetics knowledge bases

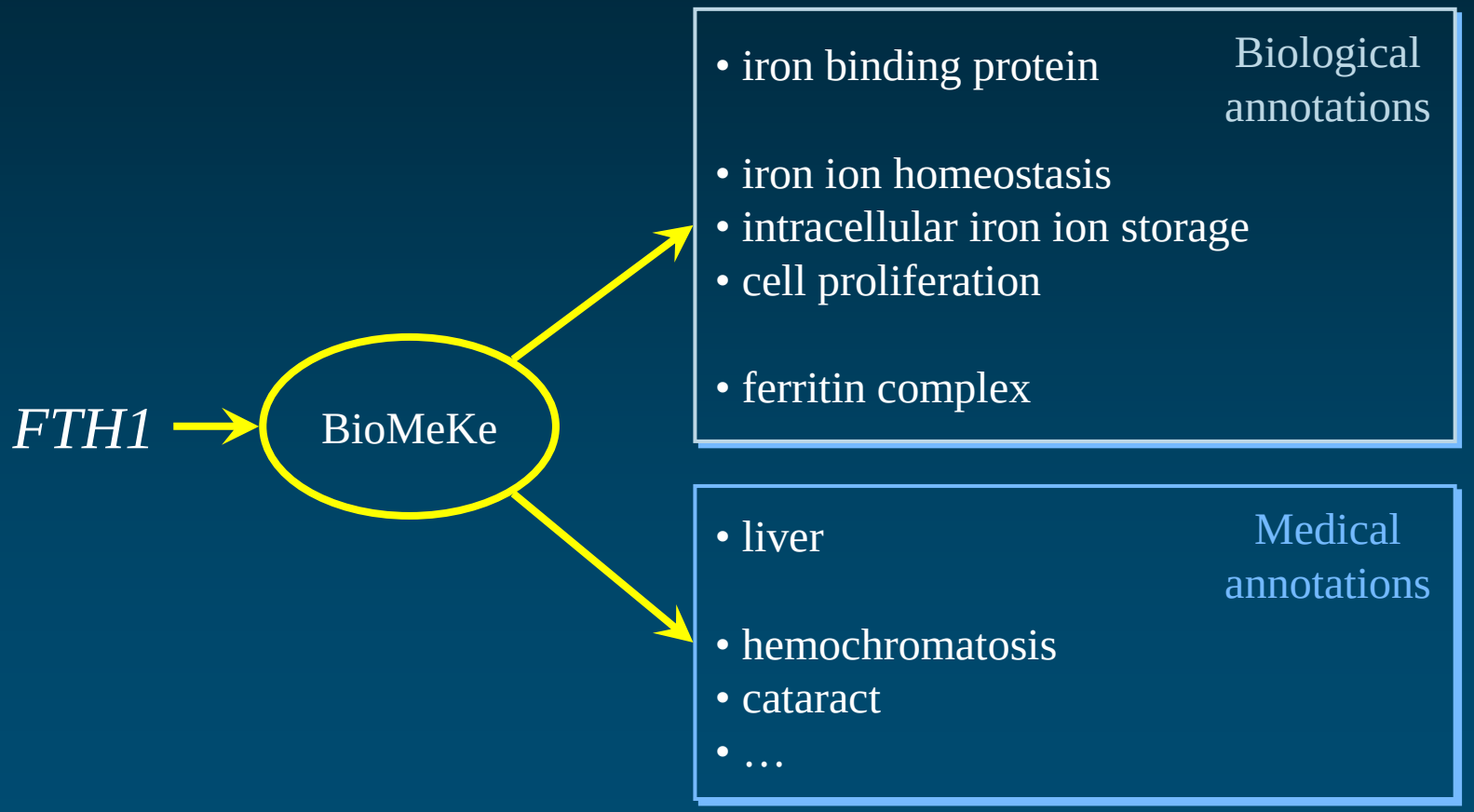
Components


Core Ontology
HUGO
UMLS
GO
Annotations
Swiss-Prot
GenBank
MEDLINE
Cross-referenced resources
Query Processor
Biological search module
Heterogeneity manager
Medical search module

Example

- Input: *ferritin, heavy polypedpide 1*
- Mapping to biological resources
  - Not found in the Core ontology
  - Official name *Ferritin heavy chain* found through Xref
- Biological information obtained from GOA
- Mapping to medical resources
  - Not found in UMLS
  - Synonym *Ferritin H* found through Xref (Swiss-Prot)
- Medical information obtained through co-occurrence of MeSH index terms in MEDLINE

Results


FTH1
BioMeKe
• iron binding protein
Biological annotations
• iron ion homeostasis
• intracellular iron ion storage
• cell proliferation
• ferritin complex
• liver
Medical annotations
• hemochromatosis
• cataract
• …

Limitations

- Non-formal ontologies
  - Knowledge may be inconsistently represented
  - Knowledge may be implicit (mappings)
- Partial automation
  - User input required to select databanks, reformulate queries
- Semantic integration
  - Naming issues
  - Mappings must be updated regularly

References

- G. Marquet, C. Golbreich, A. Burgun.
  From an ontology-based search engine towards a more flexible integration for medical and biological information.
  Semantic Integration Workshop, ISWC 2003, Sanibel, Florida, October 20, 2003;61-67.

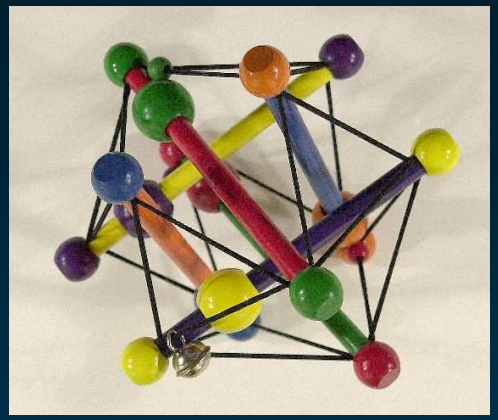
Conclusions

Conclusions

- Terminology integration provides some degree of information integration
- Most terminologies and the cross-referenced databases are readily available
- Lack of consistent representation

References

- UMLS
  umlsinfo.nlm.nih.gov
- UMLS browser
  - Knowledge Source Server: **umlsks.nlm.nih.gov**
  - Semantic Navigator: KSS, under UMLSKS resources
  - (free, but UMLS license required)
- UMLS and information integration
  - O. Bodenreider. The UMLS: Integrating biomedical terminology. *Nucl. Acids Res. 2004;32(1) (in press)*



Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: etbsun2.nlm.nih.gov:8000

NATIONAL LIBRARY OF MEDICINE

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA